Mathematical Biosciences 246 (2013) 293–304

Contents lists available at SciVerse ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

High-dimensional Bayesian parameter estimation: Case study for a model of JAK2/STAT5 signaling

S. Hug [a,b,1], A. Raue [a,c,*,1], J. Hasenauer [a], J. Bachmann [d], U. Klingmüller [d], J. Timmer [c,e,f], F.J. Theis [a,b]

[a] Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Germany
[b] Department of Mathematics, Technische Universität München, Germany
[c] Institute for Physics, University of Freiburg, Germany
[d] Systems Biology of Signal Transduction, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany
[e] BIOSS Centre for Biological Signalling Studies and Freiburg Institute for Advanced Studies (FRIAS), Freiburg, Germany
[f] Department of Clinical and Experimental Medicine, Linköping University, Sweden

ARTICLE INFO

Article history:
Available online 16 April 2013

Keywords:
Parameter estimation
Bayesian inference
Profile likelihood
Cellular signal transduction pathways
Ordinary differential equation models

ABSTRACT

In this work we present results of a detailed Bayesian parameter estimation for an analysis of ordinary differential equation models. These depend on many unknown parameters that have to be inferred from experimental data. The statistical inference in a high-dimensional parameter space is however conceptually and computationally challenging. To ensure rigorous assessment of model and prediction uncertainties we take advantage of both a profile posterior approach and Markov chain Monte Carlo sampling.

We analyzed a dynamical model of the JAK2/STAT5 signal transduction pathway that contains more than one hundred parameters. Using the profile posterior we found that the corresponding posterior distribution is bimodal. To guarantee efficient mixing in the presence of multimodal posterior distributions we applied a multi-chain sampling approach. The Bayesian parameter estimation enables the assessment of prediction uncertainties and the design of additional experiments that enhance the explanatory power of the model.

This study represents a proof of principle that detailed statistical analysis for quantitative dynamical modeling used in systems biology is feasible also in high-dimensional parameter spaces.

1. Introduction

Quantitative mathematical models can be used to describe the dynamics of cellular processes such as signal transduction. The mathematical description facilitates the understanding of complex, intertwined responses of the underlying networks of molecular reactions. The models can be employed to predict and understand features of the processes in a quantitative manner.

For building and calibrating dynamical models prior knowledge from the literature as well as quantitative experimental data is used. However, both, prior knowledge and experimental data, are limited and usually come with an associated uncertainty. These limitations and uncertainties translate to uncertainties in the mathematical model and subsequently to uncertainties of the predictions. To assess the predictive power of a model as well as its limitations, the model uncertainties have to be evaluated. As the models are often high dimensional and possess a large number of unknown parameters, this uncertainty evaluation can pose severe computational challenges. In this work we illustrate that a rigorous statistical assessment is also feasible for nonlinear high-dimensional dynamical models with over 100 parameters. For this we consider Epo-induced JAK2/STAT5 signaling, a process which has been studied extensively in recent years.

The hormone Erythropoietin (Epo) regulates erythropoesis, the production of red blood cells. Binding of Epo to its cognate receptor leads to rapid activation of JAK2 phosphorylation followed by phosphorylation of the latent transcription factor STAT5, see Fig. 1 for illustration of the model. The quantitative link between the integral STAT5 response in the nucleus and survival of erythroid progenitor cells has recently been elucidated [1]. The broad dynamical range of Epo concentrations up to 1000-fold in vivo [2] require a stringent regulatory system. In [1], it was shown that STAT5 responses are controlled by a dual feedback consisting of two inhibitory proteins, CIS and SOCS3. The two proteins adjust STAT5 phosphorylation levels over the entire range of Epo concentrations, where CIS regulates predominantly the upper concentrations range and SOCS3 the lower range. Model predictions showed that the absence (knock-out) of CIS resulted in an increase

* Corresponding author at: Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Germany.
E-mail addresses: andreas.raue@fdm.uni-freiburg.de (A. Raue), fabian.theis@helmholtz-muenchen.de (F.J. Theis).
[1] These authors contributed equally.

0025-5564/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.mbs.2013.04.002

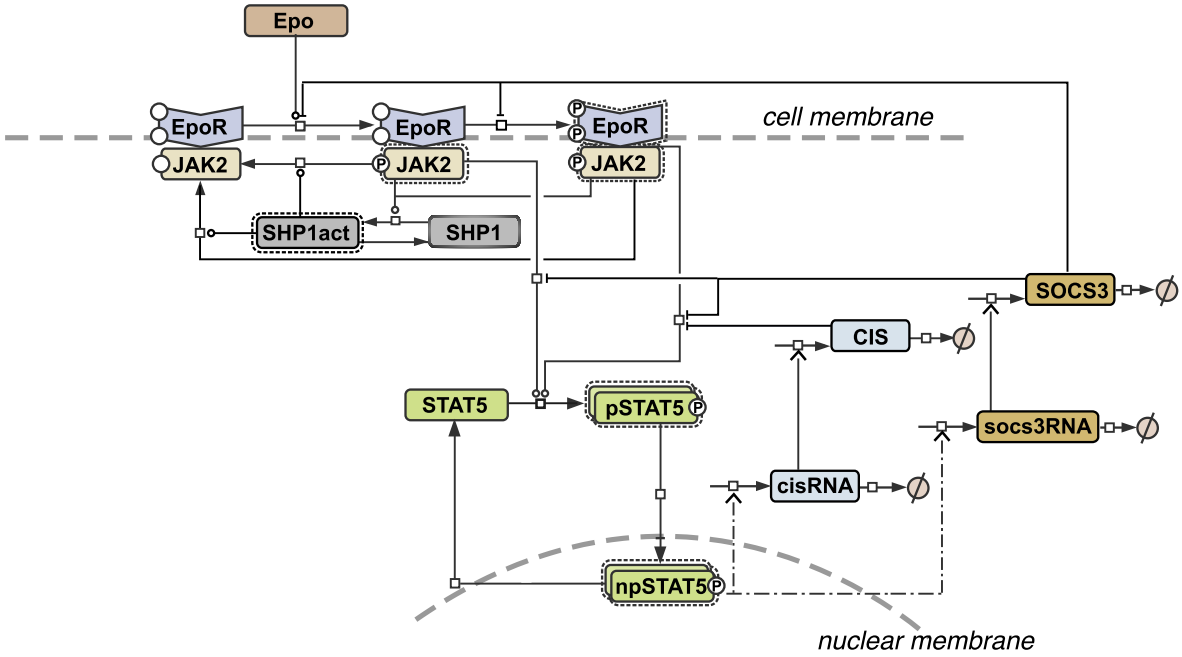

Fig. 1. Dynamical model of the Epo induced JAK2/STAT5 signal transduction pathway, adopted from [1]. The hormone Erythropoietin (Epo) binds to its membrane receptor (EpoR) and subsequently leads to receptor phosphorylation (pEpoR) and to phosphorylation of its associated Janus kinase (JAK2, pJAK2). Receptor phosphorylation is balanced by activation of a phosphatase (SHP1, SHP1act). Active EpoR/JAK2 complexes lead to phosphorylation of the Signal Transducer and Activator of Transcription (STAT5, pSTAT5) that transmits the signal to the nucleus (npSTAT5). In the nucleus, STAT5 leads to target gene expression that induces pro-survival signals and self-regulating negative feedbacks. In this case, two regulator proteins and their respective mRNAs are involved, Suppressor Of Cytokine Signaling (SOCS3) and the Cytokine-Inducible SH2-containing protein (CIS).

of STAT5 phosphorylation at low Epo concentrations, whereas the absence of SOCS3 caused an increase in the phosphorylation level at high Epo concentrations. This observation revealed division of labor by the two feedback proteins as the key property to control STAT5 responses.

In this work we analyze the Epo-induced JAK2/STAT5 signaling using a combination of maximum likelihood based parameter estimation for model calibration, identifiability analysis using the profile likelihood or profile posterior approach, and Bayesian inference using Markov chain Monte Carlo sampling (MCMC) for the translation of uncertainties to model parameters and to model predictions. Compared to the original publication [1], we carry out a refined analysis of parameter and prediction uncertainty by evaluating and interpreting posterior distributions using a Bayesian approach. The results show a second mode in the posterior that corresponds to an alternative parameterization of the model. Furthermore, the Bayesian approach reveals non-standard parameter dependency structures. The combination of high-dimensional parameter space and bimodal posterior shape represents a difficult setting for MCMC sampling. Our results indicate deficiencies of single-chain sampling schemes occurring for this high-dimensional problem with bimodal and partially flat posterior distribution. To ensure good performance for this challenging setting we instead apply a multi-chain scheme, namely Parallel Hierarchical Sampling in combination with Adaptive Metropolis Sampling, which provides improved mixing properties. Using this MCMC scheme, which is of general relevance, we can efficiently sample multimodal posterior distributions. The sampling results enable us to study the effects of the additional posterior mode on the model predictions and propose additional experiments that would allow to distinguish between both modes.

2. Methods

In this section, we present an overview of the mathematical methods applied to our biological question. First of all, this refers to modeling biological pathways as dynamical systems with ordinary differential equations. As these systems are determined by their parameters, we next present Bayesian inference as a means to inferring these parameters given the measurement data. We will then shortly introduce the profile posterior approach for identifiability analysis, which we use as a basis and complement to the Markov chain Monte Carlo (MCMC) approach. In the last section of the methods part, several of these MCMC sampling techniques which are used to sample from the complex distributions of the parameters are presented. Finally, we describe how to use the obtained samples to gain insight into the model.

2.1. Dynamical systems

Understanding cellular mechanisms has always been a key challenge of systems biology. Much effort has gone into the inference of biochemical reaction networks, such as signaling pathways, which are the main focus of this paper. All of these biochemical processes may be described by systems of biochemical reactions, where reactants are transformed into products [3]. There are various approaches for modeling the evolution of such a system over time, most common is the modeling as a dynamical system described by N ordinary differential equations (ODEs), one for each of the N modeled species, cf. e.g. [4–9]. They are characterized by a functional relationship between the current state of the system $\mathbf{x}(t) \in \mathbb{R}^N$ at time point t and its time derivative $\dot{\mathbf{x}}(t)$:

$$\dot{\mathbf{x}}(t) = f(t, \mathbf{x}(t), \mathbf{u}(t), \zeta). \tag{1}$$

This may depend on an external stimulus $\mathbf{u}(t) \in \mathbb{R}^L$ such as adding a biochemical species whose time course is not included in the model (in our case, this is e.g. stimulation of the pathway by adding Epo) as well as the dynamical parameters ζ, which are in this case e.g. the rate constants of the biochemical reactions [10]. In this approach, the dynamical parameters ζ do not depend on t and are thus constant over time. Usually, not all states of the system can actually be directly measured, so that there exists a mapping g of the